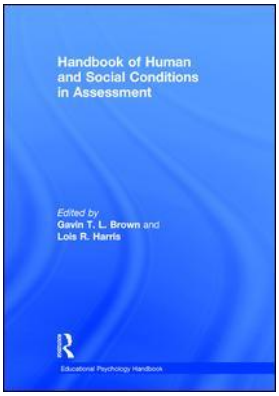

Handbook of Human and Social Conditions in Assessment

Gavin T. L. Brown, Lois R. Harris

Accountability Assessment's Effects on Teachers and Schools

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315749136.ch3
Sharon L. Nichols, Lois R. Harris
Published online on: 11 Jul 2016

3

ACCOUNTABILITY ASSESSMENT'S EFFECTS ON TEACHERS AND SCHOOLS

Sharon L. Nichols and Lois R. Harris

School accountability systems provide information to the public about how schools are performing, demonstrating that the public and/or private funds entrusted to them are leading to improved student outcomes (Cizek, 2001). Parents and taxpayers reasonably want to know if schools are doing their job. After all, businesses are held accountable, so why shouldn't schools (and those who work in them)? Despite the 'soundness' of this logic, creating a balanced and fair accountability system in education is difficult for many reasons. In business, the bottom line is easy to agree upon (profit) and measure. By contrast, in education, the desired outcomes are many (e.g., citizenship, knowledge, interest in and value for learning) and are far more difficult to measure.

While there are many potential ways to judge school success (e.g., success of graduates in the workforce, school completion rates, parent/student satisfaction surveys, or measures of student creativity), within the Anglo-American model of school accountability (see Lingard & Lewis, this volume), test scores have been promoted as a cost effective and efficient way to accomplish multiple evaluative goals. In this model, tests provide evidence of students' academic progress, but are simultaneously used to evaluate and make decisions about the effectiveness of teachers and schools. Within such an accountability system, tests are considered 'high-stakes' since there are significant consequences that can follow students' positive or negative performance, both for individual students and their teachers and schools. Following the ideology of business, where successes are judged by profit margins, here schools, teachers, and their students are judged by test scores.

Using tests to evaluate students and determine who merits particular promotions or opportunities is not new. China's *keju* system, spanning over 1,300 years (AD 605–1905), used tests to "identify and recruit the most capable and virtuous individuals into government instead of relying on members of the hereditary noble class" (Zhao, 2014, p. 32). These tests were extremely high-stakes since doing well could offer a way out of poverty and into positions of nobility and power; however, these stakes primarily affected examinees and their families. Student accountability and school accountability are two distinctly different purposes for assessment (Brown, 2008). While student accountability assessment systems (like the keju) are designed to hold the student

responsible for his or her learning, in test-based school accountability systems, a somewhat more recent phenomenon (Giordano, 2005; Herman & Haertel, 2005), results are used to determine the teacher and/or school's effectiveness and justify educational reforms.

The form and function of nations' educational accountability systems vary in two main ways. First is the locus of control. An accountability system can be directed and managed by the national government, primarily by local control, or by a mix of the two. Second, different indicators are often used; while some systems (e.g., most American states) rely primarily on one indicator (e.g., results from a battery of tests), in other systems (e.g., Finland, New Zealand) multiple measures of performance are considered (e.g., teacher observations and diverse evidence of student achievement). Accountability systems of all types may exert a level of pressure on educators to perform; however, that pressure varies enormously depending on the number and type of indicators present within the system, and the consequences attached. Undoubtedly, educators working in contexts where control is centralized and only a single indicator (e.g., test scores) is used experience the greatest pressure as the measure is narrower in focus and demands are more rigid and monolithic (i.e., the indicator is not adapted to local conditions).

The United States is an example of a country that has adopted an accountability system in which students' standardized test performance is often used as the *single* indicator of educational quality. Nationwide, students who attend public schools are subject to annual standardized tests, the results of which are often used to make decisions about the quality of their school, effectiveness of their teacher, and their academic progress. Administrators, teachers, and students are held accountable to these test scores by a system of high-stakes consequences that are triggered by students' performance. When students do well, educators are rewarded (e.g., positive publicity about the school and teacher promotion or bonuses), but if students do poorly, everyone is subject to punishing outcomes (e.g., sanctions or replacement of teaching staff). Importantly, the number and types of consequences attached to test scores varies even throughout the U.S.

The use of tests as the sole measure of educational quality is irresponsible and highly problematic. Such tests have been designed to examine what individual students know and can do, but school accountability systems use them to infer the quality of the teacher and school, usually without considering or controlling for other external factors that also influence student performance (e.g., effects of the child's previous teachers/schools; parental influences; student variables like effort during test-taking, motivation to learn at school, health and well-being). Unfortunately, humans are very quick to adopt simplistic causal explanations (e.g., scores are low, so bad teaching is to blame), even when such conclusions are not warranted (Kahneman, 2011). Those outside of the psychometric community seldom understand the limitations of assessments (e.g., measurement error), and the role of statistical processes (e.g., regression to the mean) on score increases and decreases (Gardner, 2013).

Eminent social scientist Donald Campbell warned us a long time ago about what can happen when a single indicator is used for high-stakes decision-making. He argued, "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and more apt it will be to distort and corrupt the social processes it was intended to monitor" (Campbell, 1976, p. 49). Thus, reliance on test scores, which can be viewed as a single indicator, to measure something as complicated as the process of education is likely to corrupt and distort that process, rendering the results invalid for any purpose.

It is difficult (and beyond the scope of this chapter) to examine all types of accountability systems and their effects internationally—not only do educational policies vary widely, but they change over time. Instead, we take a case study approach, reviewing what is known about the effects of educational accountability in the United States on teachers and teaching, focusing on high-stakes testing as mandated under the No Child Left Behind act of 2001 (NCLB, 2001) and the subsequent 2009 Race to the Top (RttT) grant program (U.S. Department of Education, 2009). Under these policies, the U.S. has implemented one of the most rigid, centrally imposed educational accountability systems in the world. This case explores what we have learned about the effects of these policies on teachers and serves as a cautionary tale for other countries who entertain similar ideologies of educational oversight. The discussion examines the generalizability of these findings to other contexts and provides some examples of alternative systems that may serve accountability purposes without the strong negative consequences found within the U.S. system.

A BRIEF HISTORY OF HIGH-STAKES TESTING IN THE U.S.

Virtually any test students take could be considered a high-stakes situation for the pupil, since good or bad performance has consequences (e.g., higher/lower grade point averages, tertiary entry, or job assignments). Of course, the pressures associated with these stakes vary widely depending on the type of test a student takes (e.g., college-entry vs. classroom-based) and the kind of educational goals he/she has in mind at the time (i.e., how important is this test to me?). In contrast to these everyday tests, in this chapter we use the term high-stakes testing specifically to refer to an *externally imposed* system of standardized testing where scores are used to hold administrators and teachers accountable for what students have (or have not) learned on an annual basis (Giordano, 2005; McDonnell, 2005).

The theory of action (or rationale) for high-stakes testing suggests that by tying significant consequences (i.e., incentives and punishments) to students' test score performance, teachers and their students will be motivated to work harder and more effectively, resulting in learning gains over time and a reduction in ongoing achievement gaps between student groups (Carnoy, Elmore, & Sisken, 2003; Raymond & Hanushek, 2003). In the United States under NCLB and RttT, high-stakes testing accountability is the mechanism adopted to transform and improve how schools function, teachers teach, and students learn (Jennings, 2015).

Testing and School Reform

The practice in the United States of using standardized tests to make decisions about teachers and students partly stems from decades of 'manufactured' discontent about the quality of America's public school system (Berliner & Biddle, 1995; Glass, 2008). This ongoing criticism leveled at public schools was occasionally made more salient by certain sociopolitical events. For example, Russia's 1957 satellite launch was used as evidence that the U.S. education system was rapidly falling behind, particularly in science and math (Bracey, 2008). Later, *A Nation at Risk* (National Commission for Excellence in Education, 1983) argued that the American school system was eroding rapidly, so if solutions were not found and implemented immediately, America's very economic vitality would be devastated. More recently, critics have pointed to America's average test performance on international tests (e.g., PISA) as 'proof' that it is failing to prepare students to compete globally (Armario, 2010; Berliner, Glass, and Associates,

2014; Duncan, 2010). Over time, the narrative that America's public school system is 'failing' or is in 'crisis' has helped to spur widespread support for the use of high-stakes testing (Glass, 2008; Ravitch, 2011, 2013).

Accountability Policies Featuring High-stakes Testing

For at least a century, various regions of the U.S. have experimented with high-stakes testing as a strategy for reform (Lavigne, 2014; Lavigne & Good, 2014). It was not until the early 1990s that high-stakes testing as a lever for reform was used more systematically. In the early 1990s, Texas began using standardized tests as a way to evaluate teachers and schools and to distribute significant consequences to schools (via public ranking systems), teachers (via tenure/promotion decisions), and students (diplomas were contingent on performance) (Craig, 2009). Texas school leaders and politicians touted the 'success' of this approach to school reform by holding up data that seemed to show student achievement and graduation rates were improving as a result of high-stakes testing policies. Although it was later shown that these successes were 'myths' (Haney, 2000), by 2001, politicians overwhelmingly (and unquestioningly) supported the promise of high-stakes testing accountability for transforming school outcomes through the bipartisan passage of the No Child Left Behind act (NCLB, 2001) (Nichols & Berliner, 2007a).

NCLB is a complex piece of federal legislation that made states' access to federal education tax dollars contingent on their compliance with its requirements. While containing many mandates, the central feature focused on here is the high-stakes testing component. Specifically, states had to implement five key steps to be in compliance with federal law. First, all states had to define and distribute a set of core academic standards across all grade levels and subjects. States then had to develop a battery of criterion-referenced standardized tests in core subject areas (i.e., math, reading, and science) that would gauge students' progress against these grade level standards. Next, states identified cut-score targets that would define academic proficiency levels (i.e., what score constitutes proficient vs. failing, etc.). Next, states had to develop a set of aggregate proficiency targets each school was expected to meet annually (dubbed Adequate Yearly Progress or AYP). AYP spelled out annual achievement targets that would ultimately lead to 100% of students passing the battery of tests by the year 2014. And lastly, all states had to implement a system of escalating consequences for schools if they failed to meet AYP targets. Although states varied widely in their target goals and test proficiency cut-scores, all faced a system of consequences based primarily on how students performed on the state battery of tests.

Race to the Top (RttT) was the follow-up policy to NCLB. RttT is a $4 billion competitive grants program that incentivized states to reorganize their policies to align with federal government educational reform goals. The primary feature of RttT was the use of tests to evaluate teachers. Although this was embedded throughout NCLB, with RttT the use of tests, growth scores, and other indicators to evaluate teachers became an even more central focus (Lavigne & Good, 2014). Whereas NCLB was a comprehensive approach to school reform that used high-stakes testing systems as its main platform to evaluate schools, teachers, and students, RttT encouraged states to explicitly use standardized tests to gauge teacher 'effectiveness.' It required that teachers be evaluated primarily (although not necessarily solely) on student growth (i.e., improvement in student achievement on tests over time).

Since its release, 18 states and the District of Columbia have received RttT grants. The outcome has been that states vary even more in terms of the role standardized tests

play in making high-stakes decisions about educational leaders and teachers. Although some states continue to rely primarily on students' standardized test performance for distributing consequences, in other states (such as those who received RttT grants), evaluation systems may rely on multiple indicators (test scores, and/or growth scores, and/or value added scores, and/or observation measures) and focus more centrally on using these scores to make decisions about teachers (i.e., recruitment, retention, and firing). Although states vary somewhat in terms of the number of indicators used for evaluating teachers, the core feature of high-stakes accountability testing persists and, in many states, educators continue to face higher (or increasing) pressures tied to student testing.

INTENDED AND UNINTENDED CONSEQUENCES OF HIGH-STAKES ACCOUNTABILITY TESTING

Since the mandated institution of high-stakes accountability tests in the U.S. in 2001, we have come to learn a great deal about the intended and unintended effects of high-stakes testing practices on teachers and their students within the United States (Boohrer-Jennings, 2005; Jones, Jones, & Hargrove, 2003; Nichols & Berliner, 2007a; Valenzuela, 2005). While it was the intention of the law to increase student learning and reduce ongoing and pervasive achievement gaps (as measured by tests), there is still no strong evidence either of these goals have been achieved. Additionally, data from the U.S. suggest that the NCLB and RttT high-stakes testing systems have affected preservice and in-service teachers in numerous ways.

Achievement

Isolating the impact of high-stakes testing on student achievement trends over time is challenging because these effects are often confounded with other variables (Koretz, 2008). Still, most data available seem to suggest that high-stakes testing does not improve student learning and, in some cases, may undermine it. For example, Nichols, Glass, and Berliner (2006, 2012) examined the relationship between their empirically derived measure of accountability pressure across 25 states and its relationship to student achievement as measured by the low-stakes National Assessment of Education Progress (NAEP). Their examination of NAEP trends in fourth and eighth grade math and reading revealed that high-stakes testing seemed to be positively connected to fourth grade math achievement, but was largely unconnected to eighth grade math and fourth and eighth grade reading achievement. Importantly, in some cases, high-stakes testing was inversely related to reading. Nichols et al. (2006) argued that these results suggest that high-stakes testing may incentivize greater teaching to the test (more easily done around the fourth grade curriculum), explaining why they found a positive relationship in fourth grade math only. Others have found similar results (Braun, 2004; Dee & Jacob, 2009; Rosenshine, 2003). Data suggest that high-stakes testing has not had the desired effects of increasing student achievement (Grodsky, Warren, & Kalogrides, 2009; Jennings & Bearak, 2014; Nichols, 2007; Reardon, Atteberry, Arshan, & Kurlaender, 2009; Winters, Trivitt, & Greene, 2010), reducing the achievement gap (Braun, Chapman, & Vezzu, 2010; Braun, Wang, Jenkins, & Weinbaum, 2006; Timar & Maxwell-Jolly, 2012), or increasing graduation rates (Holme, Richards, Jimerson, & Cohen, 2010; Marchant & Paulson, 2005).

There is also evidence of unintended (and largely negative) consequences from these policies. As Campbell's law suggests, when teachers are pressured to get students

to pass a particular test, undesirable effects, such as deleterious instructional practices, often result (Nichols & Berliner, 2007a; Perlstein, 2007; Ryan, 2004).

Preparing to Become a Teacher

There are connections between how students experience school and how they see education later as a teacher (Anderson, 2001; Flores & Day, 2006, Hollingsworth, 1989). For example, preservice teachers' pedagogical and instructional understandings and worldviews about learning are shaped by their previous experiences (Anderson, 2001; Hollingsworth, 1989), making it reasonable to infer that the experience of taking high-stakes tests as students might inform their practice as teachers. As Hollingsworth (1989, p. 168) suggests, our previous experiences act as "filters for processing program content and making sense of classroom contexts."

There have been few studies to examine this connection as it relates specifically to high-stakes testing contexts. One exception comes from Brown (2010), who conducted a qualitative study with eight female preservice teacher candidates from Texas—a state with a long history of high-stakes testing. Brown (2010) reported that most of the candidates believed that the tests they had taken as students were not too difficult. While initially their perceptions of teaching seemed simplistic (e.g., teachers only teach to the test), Brown (2010) found that upon probing, their understandings were actually more complex. According to one candidate, the role of the teacher is to get "the student ready for society—socially, academically, and emotionally" (p. 483). This suggests that a history of classroom experiences dictated by tests and testing do not necessarily restrict preservice teachers' views of the broader purposes of schooling.

Initial ideas about being a teacher and the role of assessments in teaching, however, are also shaped through teacher education experiences (Barnes, Fives, & Dacey, 2015). Studies suggest that preservice teachers experience conflict between their classroom observations and their learning in teacher preparation classes (Brown, 2010; Gerwin, 2004). For example, Brown (2010) found that during field experiences, teacher candidates' beliefs about what constitutes 'good' teaching grew more conflicted as they observed their cooperating teachers spending a great deal of time preparing students for the state test. Particularly salient was the issue of how to develop important skills like critical thinking, while simultaneously teaching what would be on the state test, goals often in direct conflict (Au, 2007). Given limited class time, beginning teachers may worry that excessive test preparation activities will squeeze out the more authentic, spontaneous, and critical inquiry based activities modeled as effective pedagogy in their teacher preparation program.

There are also issues relating to preservice teachers' understandings of test bias, discipline boundaries, and academic achievement. Doppen's (2007) Ohio study found that while preservice social studies teachers were able to articulate opinions about fair and unfair uses of testing, they were seldom able to identify actual bias within test items or propose viable alternatives to multiple-choice testing. These gaps made most candidates unprepared to effectively enter debates around assessment and accountability or identify what was or was not appropriate content to test within their discipline. Brown and Goldstein's (2013) study of preservice teachers revealed their confusion and uncertainty about what academic achievement actually meant. This stemmed from conflicts between their initial concept of achievement as the demonstration of academic progress versus the NCLB position that achievement only occurred when a specific proficiency target (measured solely via tests) was reached. Barrett (2009) also found that preservice and new teachers experienced internal and external conflicts

whereby internal altruistic motivations to teach were increasingly frustrated by perceptions of external control perpetrated by high-stakes testing systems. It remains unclear how preservice and new teachers may deal with these somewhat conflicting ideas about their role (i.e., complex notions around holistic education versus the pressures to teach to the test), and this area certainly warrants further research and greater attention in teacher education programs.

Being a Teacher Under High-stakes Testing

High-stakes accountability testing has significantly impacted in-service teachers within the United States; their work has intensified and expanded (Valli & Buese, 2007) and their relationships with students and instructional autonomy has been affected (Watanabe, 2007). Valli and Buese (2007) conducted a mixed methods longitudinal (2001–2005) study of approximately 150 fourth and fifth grade reading and mathematics teachers, examining how instructional practices changed as NCLB implementation began. They found teachers' autonomy and control were undermined by NCLB's curriculum demands as teachers were often expected to move through curriculum at prescribed times. Simultaneously, teachers had to interpret and use test data to provide differentiated instruction and tutoring in a strategic way to make sure underperforming students passed the test, thus increasing their workload. Tensions emerged as prescribed curriculum pacing was often at odds with the philosophy of differentiated instruction where the pace of and approach to learning is adjusted to suit the individual learner (Moon, this volume). While encouraging teachers to increase differentiation is a potentially positive effect of this testing regime, the mandate that all students be at the same place at the same time, coupled with the pressure of making sure all content on the test is 'covered,' make effective differentiation extremely difficult, especially within schools that serve disproportionately high numbers of poor and minority students (Holme et al., 2010; Orfield, Losen, Wald, & Swanson, 2004; Vasquez Heilig & Darling-Hammond, 2008).

Valli and Buese (2007) identify that changes in teachers' roles have significant consequences for their pedagogies, relationships with students, and well-being. The barrage of tests and work associated with their administration, scoring, and interpretation interfered with teachers' abilities to establish quality relationships with students. Many teachers believed the constant testing was demoralizing to students, especially those struggling: "Do we have to keep slapping it in their face?" (Valli & Buese, p. 548). In short, the emphasis on tests disrupted teachers' abilities to invest quality time with students, making tests not "worth the price of diminished relational roles with their students" (Valli & Buese, p. 548).

There are few large-scale studies to examine the connection between high-stakes testing practices and teacher work satisfaction. Although Grissom, Nicholson-Crotty, and Harrington's (2014) large-scale study examining NCLB's impacts on teacher satisfaction seemed to show little to no effect (with the exception of a small effect on perceptions of teacher cooperation), their aggregate findings likely masked the wide variation in teachers' experiences across high and low socioeconomic status (SES) contexts in the U.S. Research focusing on teachers who work in high poverty contexts has found that they experience heightened pressures as a result of high-stakes testing (Abrams, Pedulla, & Madaus, 2003; Johnston-Parsons, & Wilson, & The Teachers at Park Street Elementary, 2007; Nichols & Berliner, 2007a; Nichols et al., 2006; Pedulla, Abrams, Madaus, Russell, Ramos, & Miao, 2003; Perlstein, 2007), leading to greater teacher dissatisfaction and lowered morale (McNeil & Valenzuela, 2001; Vasquez Heilig & Darling-Hammond, 2008). Initially, accountability labels (i.e., forms of

sanctions where schools are identified as underperforming) tended to trigger teacher motivation to improve via increased effort and time, new approaches, attendance at professional development workshops, and greater collaboration to address challenges (Finnegan & Gross, 2007; Johnston-Parsons et al., 2007). However, the research suggests that over time these challenges were too great for the teachers alone to overcome (Berliner, 2013, Biddle, 2014), and they eventually became demoralized. Numerous qualitative studies (Barksdale-Ladd & Thomas, 2000; Taylor, Shepard, Kinner, & Rosenthal, 2003) illustrate the concerns teachers have relating to pressure caused by the tests (which may cause some to leave the profession) and the feeling that their autonomy and professionalism is being undermined.

Another unfortunate byproduct of high-stakes testing systems has been the creation of conditions under which some teachers turn to overt and covert forms of cheating and manipulation to help bolster their students' scores on these tests (Amrein-Beardsley, Berliner, & Rideau, 2010; Nichols & Berliner, 2007b). Data suggest those in low-performing contexts are far more likely to engage in inappropriate manipulations of test conditions or data (Jacob & Levitt, 2003). Teacher cheating may take many different forms, from telling students some questions ahead of time, allowing students more time or support than guidelines stipulate, or, in extreme cases, changing student answers on their tests. Teacher motives can vary as well. Some undoubtedly are acting to retain their jobs, while others may be concerned about the impact of continued failure on student well-being or their own promotion (Amrein-Beardsley et al. 2010). Regardless of the motivation, such manipulations undermine the validity of scores being used to make decisions about students, teachers, and schools.

The pressures of high-stakes testing throughout the U.S. have influenced teachers' instructional approaches. While Au's (2007) review of U.S. studies from 1992–2006 suggests that in 25% of studies high-stakes testing led to positive curriculum changes (e.g., broadening of curriculum, increase of collaboration and student-centered pedagogies, and integration of knowledge), the majority showed evidence of narrowing curriculum, test-centered pedagogies, and increasing fragmentation of knowledge. Au (2007) concluded that the nature of the test itself strongly influenced its effects upon curriculum. Most studies suggest the curriculum has been narrowed significantly, with tested subjects and content receiving greater time and attention (Jennings & Bearak, 2014; Vasquez Heilig, Cole, & Aguilar, 2010). Data also suggest testing-pressures may alter how certain subjects are taught. For example, in English, the pressures of testing may restrict creativity and compromise learning, subsequently undermining overall test validity (Au & Gourd, 2013). Watanabe's (2007) study of middle school English teachers in North Carolina indicated tests and test preparation took time away from broader learning objectives within the subject, undermined student motivation for reading literature, decreased collaborative activities, and made writing instruction "less like a real writer writes" (p. 335). Journell's (2010) study of social studies teachers found many were hesitant to use class time to discuss relevant current political events (e.g., the 2008 presidential election) as such discussions would take time away from preparation for a high-stakes end-of-year U.S. Constitution test. Those in school contexts where students had a history of not passing these tests were even less willing to engage with students about relevant contemporary events within their subject.

Curriculum restrictions and adjustments may be even more egregious for special education teachers who are forced to narrow the curriculum to meet the needs of the test as opposed to the specific needs of the student (Johnston-Parsons et al., 2007). An additional problem for special education teachers has to do with diagnosing and

assessing students with disabilities, as decisions about student labeling may impact what test accommodations (or exemptions) the student can receive. For example, low performing students (who are disproportionately low income, minority, and/or students for whom English is a second language) are more likely to be categorized into special education programs in high-stakes accountability environments (Artiles, 2011; Harry & Klingner, 2014). This is done to take low scorers out of the general testing pool to improve the school's average test performance, and increase the chances that the school can avoid sanctions or negative publicity (Figlio & Getzler, 2006; Jacob, 2005).

There is evidence that high-stakes testing under NCLB and RttT has created further issues in relation to the distribution of teachers and their turnover. Although it is difficult to directly connect these policies with teacher turnover given many other confounding variables, there is a large body of literature that correlates teacher turnover to the undesirable working conditions often prevalent within low SES schools (Adamson & Darling-Hammond, 2012; Loeb, Darling-Hammond, & Luczak, 2005). These already challenging working conditions have only worsened under NCLB. For example, Clotfelter, Ladd, and Vigdor (2007) found that in North Carolina, where accountability policies based on high-stakes testing had been administered since 1996–1997, teacher turnover and retention rates worsened in low performing schools. The pressures of getting students to pass tests in contexts where the challenges of teaching are already high make teachers more likely to leave (Valli & Buese, 2007).

Reback, Rockoff, and Schwartz (2011) conducted a national study of teachers' perceptions of their working conditions in schools where short-term incentive pressures to pass the state test were greatest. In these contexts, they found that teachers, especially untenured ones, reported great concerns around job security and how test scores would impact their careers. They concluded that:

> our results also raise questions concerning whether NCLB pressure motivates both tenured and untenured teachers alike, whether talented teachers are discouraged from working in schools with little chance of meeting NCLB requirements, and whether schools neglect low-stakes subjects if their performance lags far below NCLB standards.
>
> (Reback et al., 2011, p. 23)

High-stakes testing has also informed teacher placement decisions. Both Cohen-Vogel (2011) and Fuller and Ladd (2012) found that principals would strategically reassign better teachers to tested grades (where test results impacted the school) and lower quality teachers to untested grades. Principals also reported relying heavily on standardized test scores as valid indicators of which teachers to hire, to which grade levels they would be assigned, and who needed particular professional development (Cohen-Vogel, 2011). Under RttT, the use of test scores for these kinds of decisions has grown in popularity and importance since states are more strategically designing accountability systems with teacher effectiveness in mind.

Policies may also impact where teachers choose to work. Achinstein, Ogawa, and Speiglman (2004) found that a combination of teacher characteristics and backgrounds, and local and state level policy climates influenced the types of teaching conditions teachers preferred. Their mixed methods case study found that accountability pressures combined with local school management practices may lead to two groups of teachers: ones who prefer more autonomy, flexibility, and opportunity to be creative in the classroom and ones who prefer structured, scripted curriculum and direct day-to-day instructions about teaching goals. In short, the pressures of testing combined with

managerial philosophies of schools may entice specific types of teachers to particular working environments.

LESSONS FROM U.S. HIGH-STAKES TESTING ACCOUNTABILITY SYSTEMS

Data suggest that high-stakes testing under NCLB and RttT has made teaching conditions more intense and less desirable, often eroding teachers' autonomy and motivation to teach. Although there are accounts of test-based pressures initially inspiring greater teacher motivation (Johnston-Parsons et al., 2007), ongoing judgments of failure are demoralizing and undermining (Finnegan & Gross, 2007). Collectively, these factors are undoubtedly leading to higher teacher turnover, especially in lower performing schools. Additionally, there appears to be credible evidence that school and teacher actions are undermining the validity of such tests by teaching to the test, manipulating some conditions (e.g., which teachers are at which grade levels), and, in extreme cases, through dishonest behavior. These factors raise the question whether such data can be credibly used for any purpose at all.

It appears that learning to teach under accountability systems like NCLB and RttT poses a complex challenge for preservice teachers because there are often conflicts between the 'best practice' pedagogy being taught at university and the enacted practices they observe and are encouraged to take up within schools. It seems important that teacher preparation programs find a way to capitalize on these tensions to help teachers in training become mindful advocates of their future profession. All too often, new teachers internalize the expectations and norms of their school, which is problematic in schools where the primary emphasis is on passing a test. Ideally, new teachers in this context should be better prepared to counter these realities by drawing on appropriate pedagogical skills learned in their training programs, but at a minimum, teacher education programs must give preservice teachers time and space to honestly discuss these tensions and co-construct some basic strategies to help them adhere, as much as possible, to best practice pedagogy.

The flaws in the accountability system under NCLB and RttT appear twofold. First, there is an overreliance on test scores as the main indicator of student progress. As Campbell's (1976) law accurately predicts, the reality is that whenever there is overreliance on a sole indicator of performance or success, the processes it is designed to measure are likely to be corrupted, in this case via cheating, teaching to the test, and other forms of school gaming (e.g., reassignment of teachers to particular year levels). Regardless of how the corruption occurs, it ultimately means that such test scores no longer accurately measure the learning they are meant to describe, making them invalid not only for decision-making purposes, but also for any kind of diagnostic purposes that could help teachers better support learners. Here, the problem is not with the test per se; it is with the fact it is the sole (or strongest) indicator of performance. Even if testing were swapped with some other form of assessment (which would then become a sole indicator), many similar consequences would arise.

Second, the use of strong sanctions as punishment when unrealistic targets are not met further demoralizes schools that are already struggling and appears to contribute to teacher attrition. These issues pose serious problems for educational equity. With the U.S. education system (and many globally), populations are not homogenous, meaning individual schools may have differing reasons for failing to meet set performance targets. This does not mean that targets should not be set and that poor performance from low SES and minority students should be accepted as inevitable; rather, instead of

sanctions, perhaps flexible systems of support should be implemented which are long-term and driven by those who understand the unique challenges facing the particular school and students it serves.

Darling-Hammond and Rustique-Forrester (2005) noted that statewide assessments systems in the U.S. can have positive outcomes when tests are of high quality, teachers are supported and highly involved in the development and scoring of the tests, and when stakes are medium or low. When assessments are used to evaluate students and their teachers for diagnostic and/or formative evaluative purposes (and when those tests are well developed and teachers are highly supported and involved), they can offer great benefits to a society looking to improve educational processes and outcomes. But attaching high-stakes consequences to those indicators instantly increases the likelihood of corruption and distortion and renders the indicator invalid. Ultimately, it is how the test is used rather than the instrument itself that is the problem.

CONCLUSION

High-stakes testing for accountability purposes is likely to continue to thrive in societies where competition (rather than collaboration) and a business model of reform relying on bottom line outcomes are seen as the best drivers of improvement. Within educational contexts, while competition may increase motivation for some, it can be a potentially destructive mechanism, because creating winners also creates losers. Unsurprisingly, as the U.S. case illustrates, the 'losers' under NCLB and RttT are the teachers and students who work in and attend schools serving the most vulnerable. While such tests may shine a spotlight on low performing students and this may lead to increased help for some (Cizek, 2001), for schools where such students make up the majority of the population, the task of meeting benchmarks can become overwhelming, and when sanctions rather than support are offered, real academic improvement becomes unlikely (Biddle, 2014).

Although in high-stakes accountability systems, like that created by NCLB and RttT, it is easy to view 'the test' as the source of all problems, the real issue is that heavy sanctions are tied to the outcomes of interest. Adopting this kind of approach has not proved effective for creating long-term desired changes within educational settings or other workplaces, as such systems tend to undermine intrinsic motivation (Deci, Koestner, & Ryan, 1999; Kohn, 1999; Lepper, Corpus, & Iyengar, 2005). When constructed well and used responsibly, it may be appropriate to use educational assessment to identify schools and students who are not meeting requirements. However, rather than using test scores to shame, blame, or punish, they should be used to identify areas for support and improvement. In lieu of punitive accountability systems, greater social, psychological, and financial investments into our schools and our teachers are needed, alongside better programs to support the physical health and well-being of the students coming to the school. For example, in Finland (where students regularly score highly on international assessments), schools provide comprehensive support; students receive meals, health care, and psychological counseling, and teachers are highly respected (Finnish National Board of Education, 2008).

Alternatives

While it is easy to identify problems with the Anglo-American accountability model, what alternatives exist? One seemingly simple solution might be to dilute the stakes attached to accountability measures. For example, within Australia, national literacy

and numeracy testing data are collected, but without the extreme sanctions present in the U.S. system (e.g., closing schools, firing teachers); scores also have limited relationships to funding (Lingard & Lewis, this volume). However, such data are publicly reported on the MySchool website in a format that invites school comparisons by the public. Unsurprisingly, parents and the media actively use these data to compare, rank order, and label local schools. The pressure of publicly looking good has raised the stakes of this assessment and meant some undesirable practices linked to accountability (e.g., teaching to the test, trying to exempt kids from testing who may not perform well) are now being documented in Australia (Polesel, Rice, & Dulfer, 2013; Thompson, 2013). Hence, it seems that any time stakes are attached to single measures, the risks of corruption and distortion follow.

Other accountability approaches might allow for more holistic evaluations of teachers and schools. One such approach relies on school inspections or visits by an external team of evaluators. While not without its critics (Codd, 2005; Thrupp, 1998), such visits can potentially allow schools to demonstrate their progress in a richer way than test scores alone can capture. For example, in New Zealand, the Education Review Office visits schools every three years (although this cycle may be shortened or lengthened based on the school's perceived needs) (Crooks, 2011). During these visits, schools are allowed to select from a diverse range of tools (including standardized low-stakes tests and teacher judgments) to demonstrate effectiveness. Additionally, these visits examine the inputs as well as student outcomes, looking for evidence of particular practices like formative assessment; hence, evaluators can potentially make credible recommendations rather than just identifying achievement gaps. While such reports are publicly published in New Zealand on the ERO website, they are narrative and do not grade or score schools in ways which invite comparisons. However, the potential benefits of an inspection system will not occur if school evaluators overly rely on a rigid and narrow set of criteria and if criticism is not delivered in a constructive way.

Another accountability approach is to focus on the overall health of the system rather than monitor each individual school. Finland's accountability system has adopted this model; as Sahlberg (2011, p. 35) notes: "instead of test-based accountability, the Finnish system relies on the expertise and professional accountability of teachers who are knowledgeable and committed to their students and communities." Every year, a sample of schools participates in testing; however, the subject of these tests varies (e.g., mother tongue, mathematics, or other curriculum areas) and these results are not used for ranking schools; schools receive their own results back for development purposes (Ministry of Education and Culture, 2013). This light sampling process is similar to other low-stakes monitoring systems like National Education Monitoring Project (NEMP) in New Zealand and the National Assessment of Educational Progress (NAEP) system in the United States. However, in Finland this is the only external accountability measure used. While the highly competitive nature of admission into teaching (which is relatively unique to Finland) obviously contributes to public trust in teachers, this system provides food for thought for other nations as it exemplifies an alternative model of educational oversight and, importantly, one which appears to have avoided many of the educational problems associated with school accountability measures (e.g., narrowing curriculum). However, to implement a system like Finland's, there may need to be different public attitudes towards assessment than those found in the U.S., where citizens view testing as objective and believe in the comparative uses of such data (Brookhart, 2013; Buckendahl, this volume) and teachers and schools are regularly denigrated within the media (Nichols & Berliner, 2007a).

While there are clearly diverse school accountability models which can be drawn on, simply applying another jurisdiction's model is unlikely to be successful, especially where there are major differences in student compositions, or beliefs and values around educational accountability. Black and Wiliam (2005, p. 260) remind us that "the overall impact of particular assessment practices and initiatives is determined at least as much by culture and politics as it is by educational evidence and values." Clearly, there is a need for future research on the impact of culture and beliefs on the teacher uptake and public acceptance of new school accountability models that provide viable alternatives to systems which have previously relied on narrow measures.

Additionally, it is important to remember that while teachers are important change agents within schools, there are varying estimates about their actual contribution to student achievement. For example, Nye, Konstantopoulos, and Hedges (2004) attributed 7%–21% of student outcome variance to teachers, although Hattie (2003) estimated approximately 30%. Experts agree that the majority of variance is attributable to individual student and outside-the-school variables (Berliner, 2013), with socioeconomic status significantly impacting educational outcomes (Biddle, 2014). Hence, all school accountability systems must acknowledge and account for non-teacher sources of variance if trying to directly measure and reward teacher effectiveness.

Darling-Hammond (2012, p. 21) reminds us that "educational reforms based on conceptions of equity and capacity-building focusing on high-quality teaching and learning systems and access to good instruction for all students have proved to be more successful than educational reforms based on competition, incentives and sanctions." Hence, the challenge is to consider how, within the bounds of culture and social values, systems can be created within each jurisdiction which allow for transparency and the demonstration of progress, while still promoting equity. Additionally, it is important that such systems are designed to collaborate with teachers and school leaders rather than control them. It is only in systems where teachers, school leaders, and policy makers work together to provide appropriate evidence of student progress (see Lai & Schildkamp, this volume) derived from multiple measures that the corruption and distortion that Campbell's law predicts can be avoided.

REFERENCES

Abrams, L., Pedulla, J. J., & Madaus, G. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice, 42*(1), 18–28.

Achinstein, B., Ogawa, R. T., & Speiglman, A. (2004). Are we creating separate and unequal tracks of teachers? The effects of state policy, local conditions, and teacher characteristics on new teacher socialization. *American Educational Research Journal, 41*(3), 557–603.

Adamson, F., & Darling-Hammond, L. (2012). Funding disparities and the inequitable distribution of teachers: Evaluating sources and solutions. *Education Policy Analysis Archives, 20*(37). Retrieved from http://epaa.asu.edu/ojs/article/view/1053/1024

Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives, 18*(14), 1–33.

Anderson, L. M. (2001). Nine prospective teachers and their experiences in teacher education: The role of entering conceptions of teaching and learning. In B. Torff & R. J. Sternberg (Eds.), *Understanding and teaching the intuitive mind: Student and teacher learning* (pp. 187–215). Mahwah, NJ: Lawrence Erlbaum.

Armario, C. (2010, December 7). 'Wake-up call': U.S. Students trail global leaders. *Associated Press Wire*. Retrieved from tinyurl.com/prelqy6

Artiles, A. J. (2011). Toward an interdisciplinary understanding of educational equity and difference: The case of the racialization of ability. *Educational Researcher, 40*(9), 431–445.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258–267.

Au, W., & Gourd, K. (2013). Asinine assessment: Why high-stakes testing is bad for everyone, including English teachers. *English Journal, 103*(1), 14–19.

Barksdale-Ladd, M. A., & Thomas, K. F. (2000). What's at stake in high-stakes testing: Teachers and parents speak out. *Journal of Teacher Education, 51*(5), 384–397.

Barnes, N., Fives, H., & Dacey, C. M. (2015). Teachers' beliefs about assessment. In H. Fives & M. G. Gill (Eds.), *The handbook of research on teachers' beliefs* (pp. 284–300). New York : Routledge.

Barrett, B. D. (2009). No Child Left Behind and the assault on teachers' professional practices and identities. *Teaching and Teacher Education, 25*(8), 1–8.

Berliner, D. C. (2013). Effects of inequality and poverty vs. teachers and schooling on America's youth. *Teachers College Record*, *116*(1). Retrieved August 28, 2014 from http://www.tcrecord.org

Berliner, D. C., & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools.* Reading, MA: Addison-Wesley.

Berliner, D. C., & Glass, G. V., & Associates. (2014). *50 myths and lies that threaten America's public schools: The real crisis in education*. New York: Teachers College Press.

Biddle, B. J. (2014). *The unacknowledged disaster: Youth poverty and educational failure in America*. New York: Sense Publishers.

Black, P., & Wiliam, D. (2005). Lessons from around the world: how policies, politics and cultures constrain and afford assessment practices. *Curriculum Journal, 16*(2), 249–261.

Boohrer-Jennings, J. (2005). Below the bubble: 'Educational triage' and the Texas accountability system. *American Educational Research Journal, 42*(2), 231–268.

Bracey, G. W. (2008). Disastrous legacy. *Dissent, 55*(4), 80–83.

Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Educational Policy Analysis Archives*, *12*(1), 1–40. Retrieved from http://epaa.asu.edu/epaa/v12n1/

Braun, H., Chapman, L., & Vezzu, S. (2010). The Black-White achievement gap revisited. *Education Policy Analysis Archives, 18*(21). Retrieved from http://epaa.asu.edu/ojs/article/view/772

Braun, H. I., Wang, A., Jenkins, F., & Weinbaum, E. (2006). The Black-White achievement gap: Do state policies matter? *Education Policy Analysis Archives, 14*(8). Retrieved from http://epaa.asu.edu/epaa/v14n8/

Brookhart, S. M. (2013). The public understanding of assessment in educational reform in the United States. *Oxford Review of Education, 39*(1), 52–71. doi:10.1080/03054985.2013.764751

Brown, C. P. (2010). Children of reform: The impact of high-stakes education reform on preservice teachers. *Journal of Teacher Education, 61*(5), 477–491.

Brown, G.T.L. (2008). *Conceptions of assessment: Understanding what assessment means to teachers and students*. New York: Nova Science Publishers.

Brown, K. D., & Goldstein, L. S. (2013). Preservice elementary teachers' understandings of competing notions of academic achievement coexisting in post-NCLB public schools. *Teachers College Record, 115*(1), 1–37.

Campbell, D. T. (1976). Assessing the impact of planned social change. Occasional Paper Series, #8.

Carnoy, M., Elmore, R., & Sisken, L. S. (Eds.). (2003). *The new accountability: High-schools and high-stakes testing*. New York: Routledge Farmer.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19–27.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *36*(6), 673–682.

Codd, J. (2005). Teachers as 'managed professionals' in the global education industry: The New Zealand experience. *Educational Review, 57*(2), 193–206. doi:10.1080/0013191042000308369

Cohen-Vogel, L. (2011). Staffing to the test: Are today's school personnel practices evidence based? *Educational Evaluation and Policy Analysis, 33*(4), 483–505.

Craig, C. J. (2009). The contested classroom space: A decade of lived educational policy in Texas schools. *American Educational Research Journal, 46*, 1034–1059.

Crooks, T. (2011). Assessment for learning in the accountability era: New Zealand. *Studies in Educational Evaluation, 37*(1), 71–77. doi:http://dx.doi.org/10.1016/j.stueduc.2011.03.002

Darling-Hammond, L. (2012). Two futures of educational reform: What strategies will improve teaching and learning? *Schweizerische Zeitschrift für Bildungswissenschaften, 34*(1), 21–38.

Darling-Hammond, L., & Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. In J. L. Herman, & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (pp. 289–319). The 104th Yearbook of the National Society for the Study of Education (part 2). Malden, MA: Blackwell.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125, 627–668.

Dee, T., & Jacob, B. (2009, November). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*(3), 418–446.

Doppen, F. H. (2007). Pre-service social studies teachers' perceptions of high-standards, high-stakes. *International Journal of Sociology in Education, 21*(2), 18–45.

Duncan, A. (2010). Back to school: Enhancing U.S. education and competitiveness. *Foreign Affairs, 89*(6), 65–74.

Figlio, D. N., & Getzler, L. S. (2006). Accountability, ability and disability: Gaming the system? In T. J. Gronberg & D. W. Jansen (Eds.), *Improving school accountability: Check-ups or choice. Advances in applied microeconomics* (Vol. 14), pp. 35–49. New York: Elsevier.

Finnegan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's Low-performing schools. *American Educational Research Journal, 44*(3), 594–629.

Finnish National Board of Education. (2008). *Education in Finland*. Helsinki: Finnish National Board of Education.

Flores, M. A., & Day, C. (2006). Contexts which shape and reshape new teachers' identities: a multiperspective study. *Teaching and Teacher Education, 22*, 219–232.

Fuller, S. C., & Ladd, H. F. (2012, April). *School based accountability and the distribution of teacher quality among grades in elementary schools*. CALDER Working Paper, No. 75.

Gardner, J. (2013). The public understanding of error in educational assessment. *Oxford Review of Education, 39*(1), 72–92. doi:10.1080/03054985.2012.760290

Gerwin, D. (2004, March/April). Preservice teachers report the impact of high-stakes testing. *Social Studies, 95*(2), 71–74.

Giordano, G. (2005). *How testing came to dominate American schools: The history of educational assessment.* New York: Peter Lang.

Glass, G. V. (2008). *Fertilizers, pills, and magnetic strips: The fate of public education in America.* Charlotte, NC: Information Age Publishing.

Grissom, J. A., Nicholson-Crotty, S., & Harringon, J. R. (2014, December). Estimating the effects of No Child Left Behind on teachers' work environments and job attitudes. *Educational Evaluation and Policy Analysis, 36*(4), 417–436.

Grodsky, E. S., Warren, J. R., & Kalogrides, D. (2009). State high school exit examinations and NAEP long-term trends in reading and mathematics, 1971–2004. *Educational Policy, 23*, 589–614. doi:10.1177/0395909808320678

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved from http://epaa.asu.edu/epaa/v8n41/index.html.

Harry, B., & Klingner, J. (2014). *Why are so many minority students in special education? Understanding race and disability in schools.* New York & London: Teachers College Columbia University.

Hattie, J. (2003, October). *Teachers make a difference: What is the research evidence?* Paper presented at the Australian Council for Educational Research, Auckland, NZ.

Herman, J. L., & Haertel, E. H. (Eds.). (2005). *Uses and misuses of data for educational accountability and improvement. The 104th yearbook of the National Society for the Study of Education, part II*. Malden, MA: Blackwell.

Hollingsworth, S. (1989). Prior beliefs and cognitive change in learning to teach. *American Educational Research Journal, 26*, 160–189.

Holme, J. J., Richards, M. P., Jimerson, J. B., & Cohen, R. W. (2010). Assessing the effects of high school exit examinations. *Review of Educational Research*, *80*(4), 476–526. doi:10.3102/0034654310383147

Jacob, B. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*(5–6), 761–796.

Jacob, B. A., & Levitt, S. D. (2003). *Rotten apples: An investigation of the prevalence and predictors of teacher cheating.* National Bureau of Economic Research (NBER Working Paper No. 9413).

Jennings, J. J. (2015). *Presidents, congress, and the public schools: The politics of education reform*. Cambridge, MA: Harvard Education Press.

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher, 43*(8), 381–389.

Johnston-Parsons, M., & Wilson, M., & The Teachers at Park Street Elementary. (2007). *Success stories from a failing school: Teachers living under the shadow of NCLB*. Charlotte, NC: Information Age Publishing.

Jones, M. G., Jones, B., & Hargrove, T. (2003). *The unintended consequences of high-stakes testing.* Lanham, MD: Rowman & Littlefield.

Journell, W. (2010, Spring). The influence of high-stakes testing on high school teachers' willingness to incorporate current political events into the curriculum. *The High School Journal, 93*(3), 111–125.

Kahneman, D. (2011). *Thinking, fast and slow*. Sydney: Penguin Books.

Kohn, A. (1999). *Punishment by rewards*. New York: Houghton Mifflin.

Koretz, D. (2008). *Measuring up: What educational testing is really telling us*. Cambridge, MA: Harvard University Press.

Lavigne, A. L. (2014). Exploring the intended and unintended consequences of high-stakes testing evaluation on school, teachers, and students. *Teachers College Record, 116*, 1–29.

Lavigne, A. L., & Good, T. L. (2014). *Teacher and student evaluation: Moving beyond the failure of school reform*. New York: Routledge.

Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Journal of Educational Psychology, 97*(2), 184–196.

Lingard, B., & Lewis, S. (2016). Globalization of the Anglo-American approach to top-down, test-based educational accountability. In G.T.L. Brown & L. R. Harris (Eds.), *Handbook of human and social factors in assessment* (pp. 387–403). New York: Routledge.

Loeb, S., Darling-Hammond, L., & Luczak, J. (2005). How teaching conditions predict teacher turnover in California schools. *Peabody Journal of Education, 80*(3), 44–70.

Marchant, G. J., & Paulson, S. E. (2005). The relationship of high school graduation exams to graduation rates and SAT scores. *Education Policy Analysis Archives, 13*(6). Retrieved from http://epaa.asu.edu/epaa/v13n6/

McDonnell, L. (2005). Assessment and accountability from the policymaker's perspective. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement: The 104th yearbook of the National Society for the Study of Education, part II* (pp. 35–54). Malden, MA: Blackwell.

McNeil, L., & Valenzuela, A. (2001). The harmful impact of the TAAS system of testing in Texas. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 127–150). New York: Century Foundation Press.

Ministry of Education and Culture. (2013). Evaluation of education. Retrieved September 13, 2015 from http://www.minedu.fi/OPM/Koulutus/koulutuspolitiikka/koulutuksen_arviointi/?lang=en

Moon, T. R. (2016). Differentiated instruction and assessment: An approach to classroom assessment in conditions of student diversity. In G.T.L. Brown & L. R. Harris (Eds.), *Handbook of human and social factors in assessment* (pp. 284–301). New York: Routledge.

National Commission for Excellence in Education. (1983). *A Nation at risk: The imperatives for educational reform.* Washington, DC: US Department of Education, National Commission for Excellence in Education.

Nichols, S. L. (2007). High-stakes testing: Does it increase achievement? *Journal of Applied School Psychology*, *23*(2), 47–64.

Nichols, S. L., & Berliner, D. C. (2007a). *Collateral damage: How high-stakes testing corrupts America's schools.* Cambridge, MA: Harvard Education Press.

Nichols, S. L., & Berliner, D. C. (2007b). The pressure to cheat in a high-stakes testing environment. In E. M. Anderman & T. Murdock (Eds.), *Psychological perspectives on academic cheating* (pp. 289–312). New York: Elsevier.

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*(1). Retrieved July 20, 2009 from http://epaa.asu.edu/epaa/v14n1/

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2012) High-stakes testing and student achievement: Updated analyses with NAEP data. *Education Policy Analysis Archives, 20*(20). Retrieved September 16, 2012 from http://epaa.asu.edu/ojs/article/view/1048

No Child Left Behind (NCLB) Act of 2001, 20 U.S.C.A. § 6301 *et seq.* (West 2003).

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257.

Orfield, G., Losen, D., Wald, J., & Swanson, C. (2004). *Losing our future: How minority youth are being left behind by the graduation rate crisis.* Cambridge, MA: The Civil Rights Project at Harvard University.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003, March). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston, MA: Boston College, National Board on Educational Testing and Public Policy. Retrieved January 7, 2004 from http://www.bc.edu/research/nbetpp/statements/nbr2.pdf

Perlstein, L. (2007). *Tested: One American school struggles to make the grade*. New York: Henry Holt & Co.

Polesel, J., Rice, S., & Dulfer, N. (2013). The impact of high-stakes testing on curriculum and pedagogy: a teacher perspective from Australia. *Journal of Education Policy*, 1–18. doi:10.1080/02680939.2013.865082

Ravitch, D. (2011). *The death and life of the great American school system: How testing and choice are undermining education.* New York: Basic Books.

Ravitch, D. (2013). *Reign of error: The hoax of the privatization movement and the danger to America's public schools*. New York: Alfred A. Knopf.

Raymond, M. E., & Hanushek, E. A. (2003, Summer). High-stakes research. *Education Next*, pp. 48–55.

Reardon, S. F., Atteberry, A., Arshan, N., & Kurlaender, M. (2009, April 21). *Effects of the California High School Exit Exam on Student Persistence, Achievement and Graduation* (Working Paper 2009–12). Stanford, CA: Stanford University, Institute for Research on Education Policy & Practice.

Reback, R. Rockoff, J., & Schwartz, H. L. (2011). *Under pressure: Job security, resource allocation, and productivity in schools under NCLB*. Working paper 16745, National Bureau of Economic Research. Retrieved December 16, 2014 from http://www.nber.org/papers/w16745

Rosenshine, B. (2003). High-Stakes testing: Another analysis. *Education Policy Analysis Archives, 11*(24). Retrieved from http://epaa.asu.edu/epaa/v11n24/

Ryan, J. (2004). The perverse incentives of the No Child Left Behind Act. *New York University Law Review*, *79*, 932–989.

Sahlberg, P. (2011). Lessons from Finland. *The Professional Educator, Summer*, 34–38.

Taylor, G., Shepard, L., Kinner, F., & Rosenthal, J. (2003). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost* (CSE Technical Report 588). Los Angeles: University of California.

Thompson, G. (2013). NAPLAN, MySchool and Accountability: Teacher perceptions of the effects of testing. *The International Education Journal: Comparative Perspectives, 12*(2), 62–84.

Thrupp, M. (1998). Exploring the politics of blame: School inspection and its contestation in New Zealand and England. *Comparative Education, 34*(2), 195–209. doi:10.1080/03050069828270

Timar, T. B. and Maxwell-Jolly, J. (Eds.). (2012). *Narrowing the achievement gap: Perspectives and strategies for challenging times*. Cambridge, MA: Harvard Education Press.

U.S. Department of Education (2009, November). *Race to the top: Executive summary*. Washington, DC: US Department of Education. Retrieved from http://ed.gov/programs/racetothetop/executive-summary.pdf

Valenzuela, A. (Ed.). (2005). *Leaving children behind: How 'Texas-style' accountability fails Latino youth.* Albany, NY: State University of New York Press.

Valli, L., & Buese, D. (2007). The changing roles of teachers in an era of high-stakes testing accountability. *American Educational Research Journal, 44*(3), 519–558.

Vasquez Heilig, J., Cole, H., & Aguilar, A. (2010). From Dewey to No Child Left Behind: The evolution and devolution of public arts education. *Arts Education Policy Review, 111*, 136–145.

Vasquez Heilig, J., & Darling-Hammond, L. (2008). Accountability Texas-style: The progress and learning of urban minority students in a high-stakes testing context. *Educational Evaluation and Policy Analysis, 30*(2), 75–110.

Watanabe, M. (2007). Displaced teacher and state priorities in a high-stakes accountability context. *Education Policy, 21*(2), 311–368.

Winters, M. A., Trivitt, J. R., & Greene, J. P. (2010). The impact of high-stakes testing on student proficiency in low-stakes subjects: Evidence from Florida's elementary science exam. *Economics of Education Review, 29*, 138–146. doi:10.1016/j.econedurev.2009.07.004

Zhao, Y. (2014). *Who's afraid of the big bad dragon? Why China has the best (and worst) education system in the world*. New York: Jossey-Bass.